About the reliability of diagnostic statements: fundamentals about detection rates, false alarms, and technical requirements

Dipl.-Ing. Sandra Rothe,
Sebastian Felix Wirtz, M.Sc.,
Univ.-Prof. Dr.-Ing. Dirk Söffker
Chair of Dynamics and Control, University of Duisburg-Essen

1 Abstract

For modern, complex, and cost-intensive machinery and equipment automated management, system monitoring as well as integrated maintenance concepts are essential. Here, methods based on statistical learning theory are frequently proposed for classification of machine states. In this context operation- and maintenance-related decisions of machines and systems are required to be highly reliable to ensure functionality as well as safety of humans, machines, and environment. This contribution considers the problems of measuring and increasing the reliability of diagnostic statements.

Up to today, well-founded theory regarding the reliability assessment of non-destructive testing (NDT) methods is used. For instance, the Probability of Detection (POD), indicating the probability that a certain fault is detected by the particular NDT method as a function of a specific fault characteristic (i.e. crack size), is widely accepted. However, transferring the concept of POD to automated, online monitoring systems is non-trivial due to increased complexity of the decision process related to in-service applications. For instance, the effect of variable operational conditions on the reliability of a classification scheme is usually not known and, in contrast to conventional NDT, rejection of external disturbances is also not feasible. Therefore, the reliability of online-monitoring systems (including hardware and software) remains mostly unquantified.
The fundamentals about calculating commonly used rates like detection rate (accuracy), false alarm rate, and precision using training data to evaluate the reliability of a decision as well as the Receiver Operating Characteristic (ROC) are explained. Using these rates, recently developed applications for diagnosis are analyzed regarding the consequences for the decision process. The analysis of practical examples, e.g. the object (stone) detection during exposition of lignite, details the challenges of estimating the reliability of individual decisions and the possibility to increase the reliability using fusion methods, in which two or more information sources are used to define the overall decision.

The organization of this paper is as follows.

In Section 2, the commonly used performance measures regarding the evaluation of the reliability are described. In Section 3, two examples for the calculation of the performance measures and the resulting problems are explained. In Section 4, the possibilities regarding reliability improvement using fusion methods are presented. Section 5 summarizes the results and concludes this contribution, further work is suggested.

2 Introduction

The overall system reliability of complex or safety critical systems is of increasing importance. For evaluating situations, conditions, or states, classification approaches are widely used in a lot of application fields, such as fault diagnosis, image recognition, or object detection. Modern monitoring systems aim to process raw sensor data automatically using statistical learning theory (i.e. pattern recognition techniques) to obtain diagnostic statements. For further use of the information to automate the decision process regarding management, monitoring, or maintenance tasks, high detection rates and low false alarm rates are required.

In the field of Non-destructive Testing (NDT), well-founded theory has been developed to assess the reliability of testing procedures. The Probability of Detection (POD), which is a probabilistic approach to assess the reliability of an NDT method [1], is frequently used. POD curves, which describe the likelihood that a certain flaw is detected as a function of flaw characteristic \( a \) (i.e. size or depth of a crack to be detected by an NDT-approach), can be computed directly from experimental data (see Figure 1).

In the general case, a suitable decision threshold of the sensor response \( \dot{a} \) is determined. A natural lower bound of the decision threshold is given by the noise level of the measurements. As the choice of the decision threshold greatly
affects the resulting POD calculations, a suitable criterion for choosing the
decision threshold must be defined to account for the tradeoff between
minimum detectable damage size and probability of false positive detections [1].

Regarding the performance evaluation of a classifier, POD is understood as the
true positive rate [2], which is also denoted as sensitivity, recall or detection rate
(regarding object detection). However, in this case the probability of false
alarms (false positive rate) remains unquantified. Therefore, the performance
assessment of a classifier is usually based on a set of testing data with known
class labels. Here, the classifier output and true class labels are compared by
means of a confusion matrix shown in Figure 2.

From the confusion matrix, different scores, such as accuracy, precision,
specificity, recall, and false positive rate are extracted. The difference between
accuracy and recall is that accuracy considers all correct classifications in
relation to all classification assignments, recall just takes the positive
classifications (e.g. object is present) into account. Calculating the correct
positive classifications over all positive assignments, the precision value is a
measurement of the reliability of the assignments of one classifier. In general,
 improved detection rates can only be achieved at the cost of increasing false
alarm rates. The principle relationship between detection and false alarm rate is
described using the ROC curve [3], which compares the detection and false
alarm rate of a classifier. In contrast to POD the ROC-curve provides a suitable
method to assess the overall performance of a classifier [2].

![Figure 1: POD-Curve](image-url)
Furthermore, the POD curve of an NDT inspection technique is computed with respect to a fixed decision threshold using model (calibration) specimens under controlled laboratory conditions [4]. However, online diagnosis of machinery requires in service application. Here, damages evolve over time and disturbances are generally possible [5]. Consequently, the sensor output is compared to a baseline signal for damage detection, where deviations cannot be readily attributed to damage due to in-situ effects [4]. Influencing factors of NDT systems are reported as testing equipment and procedures, material and geometry of test specimens, and properties of the particular defect to be detected [1]. In contrast to this, SHM systems are affected by loading conditions [5], temperature [4], and sensor degradation [4]. Furthermore, Schubert Kabban et al. mentioned, that the assumption of independent observations is not feasible in case of continuously sampled data, because measurements performed at high acquisition rates lead to several dependent observations [6].

In the past, several ideas have been reported which address different aspects to adopt POD philosophy to SHM applications. For instance, in contrast to conventional NDT the results of SHM systems are statistically not independent due to high acquisition rates [6]. In this context, Schubert Kabban et al. proposed a new methodology to adopt POD procedures to provide compatibility with dependent measurement data, which is obtained from SHM systems [6]. Furthermore, multiple approaches developed to assess the reliability of SHM systems are summarized by Mandache et al. [4]. In particular, time-based POD is proposed to address the effect of damage evolution [4]. It is suggested to find a formulation of POD which enables stating the probability of detecting specific defect growth within a given time interval. Multi-dimensional POD is proposed to
take the effect of several in-situ effects, i.e. loading conditions, on SHM reliability into account [4]. This includes the computation of POD with respect to each influencing factor to determine the actual reliability of the SHM system in particular situations. However, the approach requires availability of quantitative information on each influencing factor. Furthermore, quantitative knowledge regarding the impact of in-situ effects on the reliability is necessary. In order to minimize the experimental effort required to determine POD, model-assisted approaches can be used [1]. Cobb et al. proposed a model-assisted approach for determining POD of crack detection in aluminum specimens using in-situ ultrasonic inspection technique [5]. Moreover, Eckstein et al. proposed a methodology to quantify SHM performance by using cumulative distribution functions to establish a probabilistic relationship between the detected and real damage size [7]. From this method, multiple metrics of SHM performance, such as minimum detectable damage size to define a lower bound of POD as accuracy of the inspection method, and probability of false alarm are derived. However, identification of the underlying distribution functions is – particularly in context of in-situ inspection techniques, where a posteriori verification of real damage size is usually not possible – still an open issue. From the aforementioned approaches to SHM reliability assessment it is noticeable, that the common weak point is characterized by missing detailed knowledge about the impact of different factors on SHM related reliability properties.

As an alternatively to the effort for reliability enhancement of suitable classifiers, the results of more than one classifier can be combined, which is denoted as information fusion. According to [8] three abstraction levels of information fusion can be divided (Figure 3): data level fusion, feature fusion, and classifier fusion (decision fusion).
The advantage of combining different classifier outputs is the independence from the application field. Once the features are selected and the classifiers are tuned, the fusion method only uses output data of the classifiers. Depending on the output data type of the classifier, classifier fusion methods are categorized in [8] and [9] into the abstract level, where the output is a single class, the rank level, where the classes are ordered by the classifier, or the measurement level, where the classifier assigns each class a confidence value. Using validation data, the abstract and rank level can be transferred to the measurement level. Commonly used fusion techniques are voting methods (e.g. weighted voting [10]), Bayesian Combination Rule [11], and Dempster-Shafer Combination [12].

3 Two examples for reliability calculation

To illustrate the problems using performance measures for evaluating reliability, two examples are given in the next chapters.

3.1 Binary classification/object detection

In this experimental example stone detection during the removal of overburden located above coal layers using bucket-wheel excavators (see Figure 4) is considered. Heavy stones falling on transport bents can cause damages to several components of the excavator which leads to higher production costs. An example for a stone to be detected can be seen in Figure 5. Measurements of acceleration sensors located at the front of the bucket-wheel are used for the detection system. Also a laser scanner located over the transportation belt is used to detect large stones lying on the belt.

Using suitable filtering and classification methods explained in [13], a dataset containing 40 samples corresponding to a stone (positive class) and 80 samples corresponding to no stone (negative class) is classified.

Figure 4: Bucket-wheel excavator [13]
In Table 1 the classification results using different sensor signals are shown. The used data contain of 40 samples with and 80 samples without effects of a stone. The performance measures detection rate and false alarm rate, which are typical for object detection as well as accuracy and precision are calculated. Depending on different sensors, the performance varies significantly. Using the laser scanner signal for object detection, the best overall performance can be achieved (high detection rate, accuracy and precision as well as low false alarm rate). Considering the acceleration sensors, an evaluation of the performance using the performance measurements is difficult. Regarding the accuracy, all of the results using acceleration sensors are in a similar range (around 75 %). But considering the detection rate as well as the false alarm rate, they significantly differ. Using the measurements from sensor 5, a contradiction between different performance measures appear. The detection rate is the worst considering the given results, but precision is the highest among the acceleration sensors.

Table 1: Classification results of stone detection

<table>
<thead>
<tr>
<th></th>
<th>Acceleration sensor</th>
<th>Laser scanner</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Objects detected</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td>False alarms</td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>Detection rate [%]</td>
<td>55,00</td>
<td>50,00</td>
</tr>
<tr>
<td>False alarm rate [%]</td>
<td>11,25</td>
<td>16,25</td>
</tr>
<tr>
<td>Accuracy [%]</td>
<td>80,07</td>
<td>74,00</td>
</tr>
<tr>
<td>Precision [%]</td>
<td>70,97</td>
<td>60,61</td>
</tr>
<tr>
<td>F-Score [%]</td>
<td>61,97</td>
<td>54,80</td>
</tr>
</tbody>
</table>
This means, that if the decision using measurements from sensor 5 is positive, this assignment is more reliable than the positive assignments using the other sensors. Therefore a performance measure denoted as F-score can be used to combine the detection rate (recall) with precision value.

### 3.2 Fault detection for composite materials

Composite materials are increasingly used instead of conventional construction materials due to advantageous mechanical properties, such as high specific strength resulting from their complex structure. However, the more widespread use is currently restricted by the lack of ductility compared to metallic materials as well as several micromechanical damage modes [14]. Therefore, different NDT techniques, which indicate developing damages, are proposed for continuous monitoring of these materials to ensure equal degree of reliability and safety [15]. This section is based on the work presented in [16]. To assess the reliability automatic damage detection techniques, a showcase algorithm for damage classification in composites has been implemented using Acoustic Emission (AE) measurements and Support Vector Machine (SVM).

The AE technique exploits the phenomenon of elastic waves emerging from localized sources within structures or materials on damage initiation or propagation due to external loads. These surface waves propagate through structures over a significant distance and can be detected with suitable sensor equipment. Frequently, time domain features, i.e. count rate, rise time, or duration are used to discriminate different source mechanisms. This approach requires a set of empirical parameters (i.e. amplitude threshold, timing parameters) which are strongly dependent on experimental conditions and therefore need to be set by the operator. Regarding composites, the frequency content of AE waveforms is considered a more reliable descriptor of underlying source mechanisms [17]. Therefore, peak frequency analysis was used by several researchers to identify characteristic frequencies of distinct AE source mechanisms related to different failure modes of composites [18].

In [16], experiments were conducted using a bending test rig to subject specimens of composite material to cyclic bending load with different amplitudes and frequencies. Damages within the material are detected by means of SVM based classification of AE measurements using frequency-domain features. Measurements were performed using 5 different excitation frequencies in the range of [2 Hz 6 Hz] and 5 different amplitudes in the range [6 mm 18 mm]. Accordingly, the measurement series of each specimen comprises a total of 25 datasets. The classification performance is assessed using probability estimation indicating the likelihood that a particular observation is actually a member of the assigned class.
From the classification results, mean values of probability estimation were computed for dataset to investigate dependences between classification reliability and loading conditions. The results of two independent specimens S-I and S-II, which were subjected to identical testing procedure, are presented in Figure 6. Here, debonding is chosen as an example. Corresponding results considering additionally delamination, matrix crack and fiber breakage are presented in [16]. The loading pattern is plotted on the x- and y-axes, respectively. Corresponding probability estimates are indicated by the color scale. The experimental results show that the classification performance varies strongly with the excitation. In both cases, a trend of increasing probability estimates with increasing load intensity is observed. Furthermore, best results are obtained at similar excitations. However, strong differences in the performance of the damage detection algorithm are noticeable between both measurement series. Whereas, debonding is only detected at excitation frequencies below 4 Hz in case of S-I, this damage mode is also detected above 4 Hz in case of S-II.

The experimental results indicate strong dependence of (i) damage detectability and (ii) classification reliability on loading conditions. Nevertheless a direct relationship between different loading patterns and classification reliability cannot be established due to large scattering of the results. Similar results were previously reported by Gagar et al. [19]. Therefore, the question raises, which method is feasible to obtain reliable statements regarding the current state-of-health of technical systems. Several approaches to adopt POD philosophy to SHM applications are summarized by Mandache et al. [4] and Eckstein et al. [7]. However, these approaches require detailed knowledge regarding the impact of different influencing factors on the reliability of the SHM system. This
information is usually not available and – particularly in context of in-situ inspection techniques, where a posteriori verification of the results is in the general case not possible – non-trivial to obtain.

4 Improvement of reliability using fusion methods

From the literature contributions, like [10], with the focus of improving the performance of the decision support system by using simple fusion methods like voting methods (Weighted Voting (WV)), sum rules, or averaging probabilities for diagnostics of an accelerometer [10] are known. In the application field of fault diagnosis, the improvement of feature selection or classification approaches is a challenging problem. In [11] classification results are combined using the Bayesian belief method (or Bayesian Combination Rule (BCR)), which is a fusion method on measurement level based on conditional probability. In the cited contribution the goal is to reach higher accuracy of classification results for fault diagnosis of gearbox and locomotive bearings. The Dempster-Shafer Combination (DSC) based on the Dempster-Shafer Theory is another fusion method on measurement level. For fault diagnosis, this method can be used to combine neural network classifier outputs as given in [12] for fault diagnosis of induction motor, of railway track circuits [20], or for spark plug fault recognition applied on different data sources [21]. Resulting from the study in [22], the DSC is able to achieve higher probability of detection than individual classifiers. In [23] several classifier fusion methods are compared to analyze the performance to Structural health monitoring systems. First, nine fusion methods with different combinations of classifiers are compared using synthetic data. The DSC and the fuzzy logic type 2 (FLT2) algorithms were claimed to have best results concerning the correct classification rate. To validate the results, DSC and FLT2 are compared to fuzzy logic type 1 and majority voting using experimental data of structural damages of an aluminum plate. The applied fusion methods can improve the accuracy of a decision support system.

Using the example of stone detection from chapter 3.1 the results using the fusion methods BCR and an extension of the DSC with WV are compared. The results for detection rate and false alarm rate for the same dataset used in chapter 3.1 are shown in Table 2. Combining the results of the acceleration sensors, the detection rate cannot be improved, but the false alarm rate is reduced compared to the individual results with the same detection rate. The fusion of all sensors shows that the BCR and DSC with WV are working different.
Table 2: Fusion results of stone detection

<table>
<thead>
<tr>
<th></th>
<th>BCR</th>
<th>DST + WV</th>
<th>BCR</th>
<th>DST + WV</th>
<th>BCR</th>
<th>DST + WV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fusion of results based</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>on acceleration sensors</td>
<td>20</td>
<td>20</td>
<td>37</td>
<td>26</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Fusion of all results</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fusion of combined results based on acceleration sensors with result based on laser scanner</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Objects detected</td>
<td>20</td>
<td>20</td>
<td>37</td>
<td>26</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>False alarms</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Detection rate [%]</td>
<td>50,00</td>
<td>50,00</td>
<td>92,50</td>
<td>65,00</td>
<td>90,00</td>
<td>90,00</td>
</tr>
<tr>
<td>False alarm rate [%]</td>
<td>5,00</td>
<td>5,00</td>
<td>3,75</td>
<td>1,25</td>
<td>2,50</td>
<td>2,50</td>
</tr>
</tbody>
</table>

While DSC with weighted voting gives more weight to the majority of votes, BCR takes the classifiers with better performance (here laser scanner) more into account. The fusion of all results using BCR results in the highest detection rate, using DST with WV results in the lowest false alarm rate.
For further evaluation of the results the ROC can be used as shown in Figure 7, to compare both detection and false alarm rate. The optimum is at a detection rate of 100 % and a false alarm rate of 0 %. The closer results obtained to this point the better. From Figure 7 it can be concluded that, because of the least distance to the optimal point, the fusion of all sensors using BCR leads to the best performance with respect to equally weighted detection and false alarm rate, even better than the best individual result (laser scanner). Application depending users have to decide if detection rate maximization or false alarm minimization has to be preferred. The fusion of results can be used to improve the performance of a detection system without the necessity to improve every individual classification system.

5 Summary and conclusion

In this contribution the fundamentals about performance measures evaluating the reliability of detection systems are described and analyzed. Using two different applications, the problems regarding the commonly used performance measures are shown. Although there are several performance measures, the meaning of them differs and sometimes contradicts. The example of stone detection shows, that one classification system can have a high accuracy and precision value, but also low detection rate. Also the objectives of detection rate maximization and false alarm minimization conflict with one another. So the reliability evaluation strongly depends on the application depending use of the detection system. Many other impact factors are influencing the performance measures, like the POD-calculation of debonding of composite materials. The results show that the classification performance varies with different loading conditions. In general the application of performance measures to SHM systems requires detailed knowledge regarding the impact of several influencing factors like operating conditions, load conditions, etc. Also knowledge about the signal behaviors related to specific physical effects is required and the training data has to cover all these effects. This knowledge is often not available or not easy to obtain.

Using fusion of results, there is no need to develop one high reliable classifier or detection system, because the performance can be improved by combining assignments of classifiers with different performance. Some influencing factors just influencing only one individual classification can be compensated.

For future works the reliability definition as well as the performance measures for SHM-Systems should be standardized. Also the development of fusion
method and the application to SHM-Systems are within the research focus to ensure a high reliability.

6 References


